# Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships


## Authors:

Florian Huber[1], Lars Ridder[1], Simon Rogers[2], Justin J.J. van der Hooft[3]

Affiliations:

[1]Netherlands eScience Center, Amsterdam, the Netherlands
[2]School of Computing Science, University of Glasgow, Glasgow, United Kingdom
[3]Bioinformatics Group, Plant Sciences Group, University of Wageningen, Wageningen, the Netherlands



## Abstract

Spectral similarity is used as a proxy for structural similarity in many tandem mass spectrometry (MS/MS) based metabolomics analyses, such as library matching and molecular networking. This is based upon the assumption that spectral similarity is a good proxy for structural similarity. Although weaknesses in the relationship between common spectral similarity scores and the true structural similarities have been pointed out, little development of alternative scores has been undertaken. Here, we introduce Spec2Vec, a novel spectral similarity score inspired by a natural language processing algorithm -- Word2Vec. Where Word2Vec learns relationships between words in sentences, Spec2Vec does so for mass fragments and neutral losses in MS/MS spectra. The spectral similarity score is based on spectral embeddings learnt from the fragmental relationships within a large set of spectral data. Using a dataset derived from GNPS MS/MS libraries including spectra for nearly 13,000 unique molecules, we show how Spec2Vec scores are more proportional to structural similarity of molecules than the commonly used cosine score and its derivative, the modified cosine score. We also demonstrate the advantages of Spec2Vec in library searching for both exact matches and analogues as well as in molecular networking. Finally, Spec2Vec is also computationally more scalable allowing us to search for structural analogues in a large database within seconds.


## Significance Statement

Most metabolomics analyses rely upon matching observed spectra to library spectra for structural annotation. Scoring functions are used to assess the similarity between pairs of spectra. No studies have so far proposed scores fundamentally different to the popular

cosine-based similarity score, despite the fact that its limitations are well understood. We propose a novel spectral similarity score known as Spec2Vec that i) correlates better to structural similarity than cosine-based scores, ii) performs better in library matching tasks, and iii) is computationally more scalable than cosine-based scores. Given the central place of similarity scoring in key metabolomics analysis tasks such as library matching and molecular networking, we expect Spec2Vec to make a broad impact in all fields using untargeted metabolomics.

# Main text:

## Introduction

Metabolomics, the high throughput characterisation of metabolites present in a biological sample, is increasingly important across the biomedical and life sciences [1,2]. This is largely due to the manner in which the metabolome complements the genome, transcriptome and proteome as the data type most closely representing phenotype[3], and to the increased sensitivity and coverage of modern measurement equipment.

Of the available measurement platforms, liquid chromatography coupled to mass spectrometry (LC/MS) is the most widely used. Modern untargeted LC/MS experiments produce large datasets that are challenging to fully analyse and exploit. A main bottleneck is the structural annotation and identification of the chemical ions detected by the mass spectrometer, primarily because the mass-to-charge ratio (m/z) of an observed ion species is very often insufficient to unambiguously assign it to one chemical formula, and certainly insufficient to assign it to a specific chemical structure. Building upon the assumption that molecules fragment in a manner that is dependent on their structure, fragment (also known as MS2 or MS/MS) data is often used to overcome this bottleneck. Fragmentation spectra can act as an aid to annotation via either comparison with databases [4] such as MassBank, Metlin, GNPS, etc, or as input to in-silico identification algorithms such as SIRIUS/CSI:Finger ID [5] or MS2LDA [6].

At the heart of much analysis of mass spectrometry fragmentation data is the computation of similarity between pairs of MS2 spectra (Fig.1 A,B). For example, when searching an unknown spectrum against a database, a similarity score is computed between the query spectrum and database spectra. Similarly, when creating molecular networks [7], edges are drawn between spectra if their similarity exceeds a user-defined threshold. In all such analyses, spectral similarity is being used as a proxy for structural similarity, the real quantity of interest [8]. Cosine-based scores are the most widely used measures of spectral similarity. Studies investigating whether molecules with high structural similarity result in spectra with a high spectral cosine similarity score only partly support the assumed relationship between spectral and structural similarity [9,10]. As a result, various modifications of the cosine similarity score have been proposed, including raising the m/z and intensity components to different powers,

and shifting fragment peaks by the difference in precursor m/z ('modified cosine score' [7]). Cosine-based methods are very good at revealing identical or nearly identical spectra, but by conception they are not well-suited to handle molecules with multiple local chemical modifications. In addition, most cosine scores are computed by aligning the fragment peaks from the two spectra, a computationally intensive procedure that makes extensive library searching very slow. Despite these limitations, thus far no fundamentally different spectral similarity scores have been proposed.

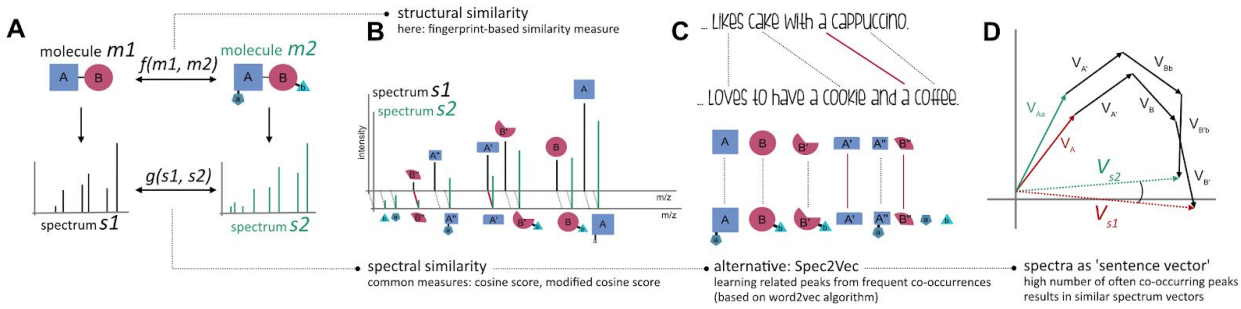



**Figure 1**. (A) MS-MS spectra can be considered as signatures of molecules: spectra are known to contain structural information of the original molecule, but without a straightforward way to translate mass spectral features into structural ones describing the fragmented molecule. (B) Spectra are commonly compared by similarity measures such as cosine or modified cosine scores. While those measures are very good at revealing (nearly-) identical spectra, they often underperform when it comes to spectra of complex molecules which are different yet similar. (C) Spec2Vec is based on algorithms from natural language processing and learns relationships between peaks which are not identical, but often co-occur. (D) Spectra from different yet similar molecules which contain many non-identical but still related peaks will be represented by similar spectral vectors.

Within the context of metabolite identification from MS/MS data, much effort has gone into the prediction of structural information from spectra. This is because it allows spectra to be queried against structural databases that are typically orders of magnitude larger than spectral ones. Despite the fact that database searching is still the gold standard for metabolite annotation[11], methods such as MAGMa[12], SIRIUS[13], CSI:FingerID[5], IOKR[14], DeepMASS[15], and MetDNA[16] have been effective at widening the search space and clearly demonstrate that useful structural information can be learnt from MS2 spectra.

Building upon the insight that structural information can readily be learnt from spectra, we present a novel spectral similarity score based upon learnt embeddings of spectra. Inspired by the success of algorithms from the field of natural language processing in accounting for element relatedness for overall similarity assessment of objects, we aimed at adapting such tools to mass spectra data. A language-based analogy would be words like 'cookie' and 'cake' which often occur in similar contexts (e.g. together with words like 'dough', 'sweet', 'eating') and are hence assumed by the model to represent very related entities (Fig.1 C).

Adapting Word2Vec, a well-established technique in natural language processing [17], Spec2Vec learns meaningful vector representations of peaks and spectra using their m/z values and intensities. Similarly, Spec2Vec learns from co-occurrences across large datasets to represent

highly related fragments or neutral losses by vectors pointing in similar directions within a continuous latent space. A spectrum can then be represented by a low-dimensional vector calculated as the weighted sum of all its fragment (and loss) vectors (Fig.1 D). Instead of relying on only a binary assessment of each fragment (match/no match), Spec2Vec hence takes the relatedness between fragments into account. This can be illustrated by comparing two spectra of molecules with multiple (small) chemical modifications that Spec2Vec could relate to each other despite the low amount of direct peak matches (Fig. 2).

In contrast to the machine learning methods mentioned above, Spec2Vec is unsupervised and can be trained on any collection of spectra. We demonstrate that Spec2Vec similarity scores are well suited to identify structural similarities on the basis of given MS/MS spectra. Spec2Vec similarity scores are very fast to compute, which -taken together- creates promising use cases of Spec2Vec in library matching as well as molecular networking.

# Results

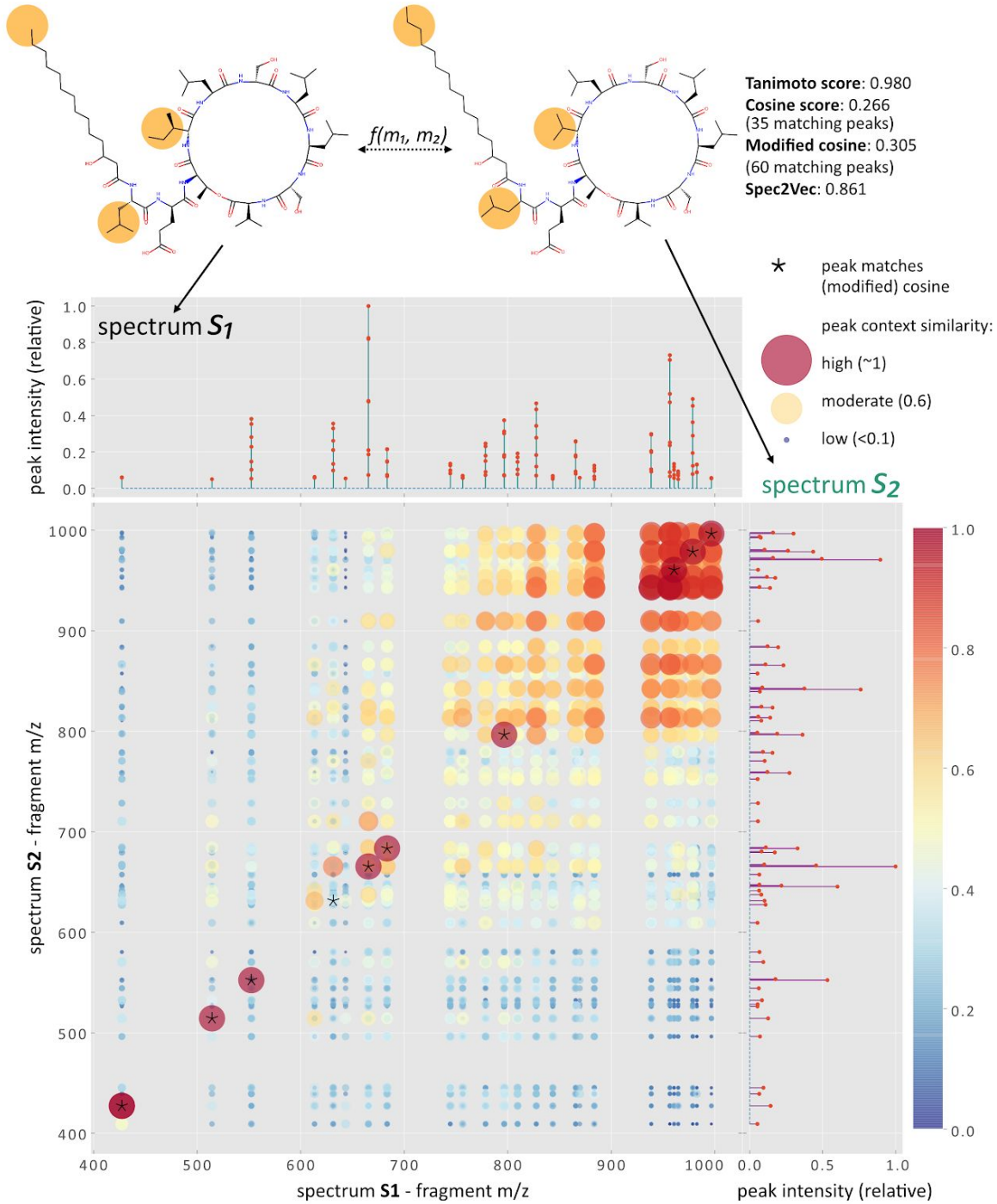



**Figure 2**. In-depth comparison example of two spectra. Since the two molecules differ slightly in three locations, both cosine and modified cosine scores fail to recognize the overall structural similarity and return low spectral similarity scores. Spec2vec for many peaks acknowledges that they often co-occur across the training data, hence showing a high peak context similarity which overall leads to a high Spec2Vec similarity score. For illustrative purposes, this figure only displays peaks between 400 and 1000 Da.

## Spectra similarity vs structural similarity

To quantify how well different spectral similarity scores correlate with structural similarity, we calculated multiple spectral similarity scores for all possible pairs across a mass spectra dataset of representative character, but computationally manageable size. To create this dataset, we started with a large collection of mass spectra provided by GNPS. Library spectra were used to allow us to compare spectral similarity scores with scores based on (the known) molecular structures. It is important to note, however, that Spec2Vec is an unsupervised machine learning technique that can be trained on any collection of spectra, independent of the presence of annotations. After filtering, processing, and removing all spectra with fewer than 10 fragment peaks (see methods), the remaining **AllPositive** dataset comprised 95,320 positive ionization mode mass spectra, 77,092 of which had inchikey annotations. Because many similarity scores are computationally expensive, the quantitative similarity score assessment was done on a subset of this data, **UniqueInchikey,** consisting of 12,797 spectra with unique Inchikeys (first 14 characters thus planar InchiKeys, see methods).

For the **UniqueInchikey** data it was possible to compare the different spectra similarity scores to the structural similarity, represented by Tanimoto scores. One of our core interests was to evaluate to what extent a high spectral similarity score, $g(s_1, s_2)$ between a pair of spectra reflects a high structural similarity score $f(m_1, m_2)$ between the respective molecules . For this, we computed different similarity scores for all possible spectra pairs, hence between all 12,797 spectra in **UniqueInchikey**. The vast majority of those spectra pairs correspond to entirely unrelated molecules, resulting in a distribution of fingerprint-based structural similarities with a very strong shift to low Tanimoto scores as shown in fig. 3A. We then selected only the structural similarities that correspond to the highest scoring pairs according to each similarity measure (Spec2Vec, Cosine, modified Cosine). Figure 3B displays the average structural similarity over the highest 0.1% of each respective spectra similarity score (0.1% here stand for about 80,000 spectra pairs). This reveals that a high Spec2Vec spectrum similarity score correlates stronger with structural similarity than the cosine or modified cosine scores (Fig.3). Cosine scores exist in numerous flavors and can vary largely depending on their key parameters (here mostly: tolerance and min_match, the minimum number of matching peaks). Several different cosine score flavors and parameter ranges were tested, without resulting in major improvements regarding figure 3 (see supplemental material and figures S2 to S6).

We observed that the poorer correlation between cosine and modified cosine similarity scores and structural similarity can largely be explained by high false positive rates (fig. S1). This shortcoming can to some extent -- though never fully -- be reduced by using lower peak match tolerances (here: tolerance = 0.005 Da) and ignoring scores based on fewer than *min_match* matching peaks (here: *min_match*=10, see fig. S2, S3 for parameter search).

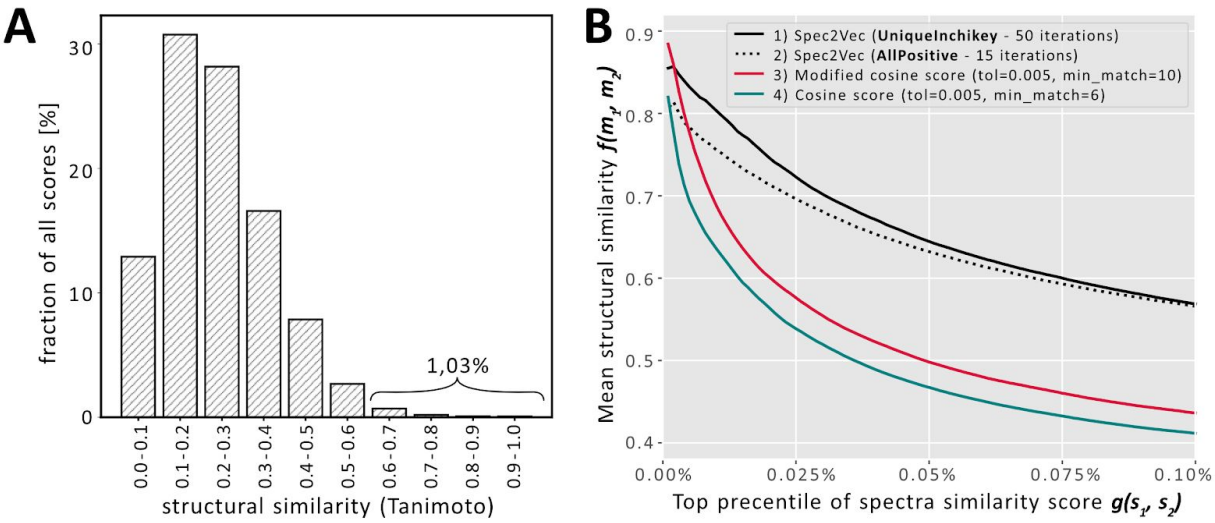



**Figure 3.** (A) histogram of the structural similarity scores across all possible spectra pairs between the 12,797 spectra in the **UniqueInchikey** dataset (81,875,206 unique pairs, not including pairs of spectra with themselves). The histogram indicates that randomly chosen pairs will most likely show scores between 0 and 0.5. Similarity scores > 0.6 are very infrequent and hence unlikely to achieve by randomly choosing pairs (p=0.0103 of finding a score > 0.6, p=0.0034 for a score > 0.7). (B) Different similarity scores were calculated for the same 81,875,206 spectral pairs. Comparing the highest 0.1% the resulting scores to the structural similarities reveals that Spec2Vec similarities show a notably higher correlation with actual structural similarities.

## Library matching.

Next, we evaluated the potential of Spec2Vec to aid in matching unknown spectra to library spectra run on various instruments under different conditions. We here worked with the **AllPositive** dataset (95,320 spectra), of which we removed 1000 randomly selected spectra such that they would all have a different planar InchiKey and that we had at least 4 spectra with identical InchiKey remaining. The remaining 94,320 spectra served as the library and are hereafter referred to as 'database'. Therefore, for each of the 1000 held-out spectra, multiple positive hits existed in the database. A Spec2Vec model was trained on the database and Spec2Vec scores were compared with cosine similarity scores for library matching (Fig.4). Both Spec2Vec and cosine similarity scores were used in the same way: potentially matching spectra were pre-selected based on precursor mass matches (tolerance = 1Da) before the highest scoring candidate above a similarity threshold was chosen. Gradually lowering this threshold from 0.95 to 0, increases both the number of true and false positives per query. While the general trend for both scores is similar, Spec2Vec resulted in a notably better true/false positive ratio at all thresholds. Spec2Vec also allowed to correctly match the query spectra with up to 87% accuracy and showed both higher accuracy and retrieval rates when compared to the cosine score based library matching (Fig. 4).

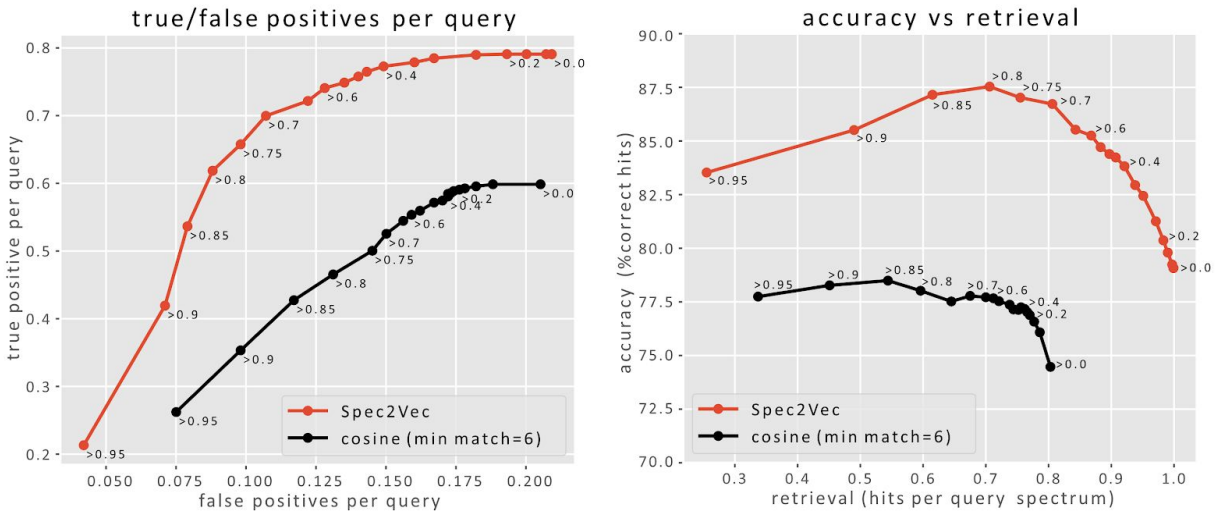



**Figure 4**. Spec2Vec similarity scores deliver improved true-to-false-positive ratios during library matching. 1000 randomly selected spectra, all with at least 5 identical InchiKeys in the entire dataset, were removed from a **AllPositive** and then matched to the remaining spectra. Matching was done by pre-selecting spectra with the same parent mass (tolerance = 1Da) and then choosing the candidate with the highest spectral similarity score if this score was larger than a set threshold. The left plot shows the true-vs-false ratio per query when using Spec2Vec (red) or Cosine scores (black). Labels near the dots report the used similarity score thresholds. The plot on the right displays the resulting accuracy and retrieval rates for the same parameters. Using Spec2Vec, library matching could be done with notably higher accuracy across all tested retrieval rates.

In practice, we expect that actual library matching can be improved further via stricter parent mass matching (lower tolerance) or by consulting both Spec2Vec and Cosine similarity scores, which can potentially make use of the apparent complementarity of the methods (as described in the next section).

## Unknown compound matching

Once the embedding has been trained, Spec2Vec similarity scores are computationally more efficient to calculate than cosine-based scores, allowing brute-force all-vs-all comparisons of query spectra against large datasets. Although computation times for the different similarity scores will depend upon multiple factors such as the chosen spectra preprocessing steps (key variable: resulting number of peaks), and implementation or hardware details, we generally found Spec2Vec similarities to be about 2 orders of magnitude faster to calculate when using a pre-trained model, and about 1 order of magnitude faster when training a new model from scratch. Training a model on the UniqueInchikey dataset takes about 30 minutes on a conventional CPU (here: Intel i7-8550U). The main performance gain results from the fact that Spec2Vec embedding vectors are fixed length vectors, comparison of which is straightforward. Conversely, cosine-based methods, despite their name, do not tend to operate on fixed vectors as this would require a binning step. Instead, they rely on a costly alignment step in which all pairs of fragment ions with matching m/z (within some tolerance) are extracted. In addition, Spec2Vec similarity correlates more closely with structural similarity (fig.3) which makes it

better suited for identifying relationships between spectra of different yet chemically related molecules. Taken together, this makes it possible to use Spec2Vec similarity for rapidly querying spectra of unknown molecules against all spectra in a large database (fig.5).

To specifically test if Spec2Vec similarity scores can be used to help detect highly related molecules, we moved all spectra belonging to 200 randomly chosen IchiKeys (1030 spectra in total) from the **AllPositive** dataset to a query dataset. To make sure that there is no overlap regarding molecules between training and query data, we also removed all spectra without InchiKey annotation. A separate word2vec model was trained on the remaining data (76,062 spectra), and therefore the model did not see any spectra of the selected 200 molecules. Finally, each of the 1030 query spectra was compared to all remaining spectra. The ten highest scoring matches for each query were selected and the quality of this selection was assessed by computing the structural similarities between all found matches and the query molecules (fig.5). For 3 out of 5 queries, the found top-10 list would contain suggested molecules with a significantly higher than chance structural similarity score of > 0.6 (p=0.0103, see fig. 3A for histogram of scores and calculation of the percentage). In particular for molecules with parent masses larger than 200, Spec2Vec was often able to detect very related molecules for unknown queries. For query spectra of larger molecules (>400 Da), Spec2Vec similarities were very likely to point to chemically highly related library molecules with an average Tanimoto similarity for the best suggestions above 0.8 (fig.5 B).

This experiment on purpose only considered Spec2Vec as the only means to find related matches to best illustrate its ability to quickly find related molecules in a large library dataset. Querying 1030 spectra against 76,062 spectra took 110s on a standard CPU (here: Intel i7-8550U), or 0.1s per query spectrum. This makes Spec2Vec similarity scores a very powerful selection tool. With a total of $1,030 \cdot 76,062 \approx 78 \cdot 10^6$ similarity scores to calculate, we estimate the same operation using cosine or modified cosine would take about 4-8 hours, which is the time it took to compute the $82 \cdot 10^6$ similarity scores for the benchmarking comparison (fig. 3). The high computational cost of the cosine and modified cosine scores is also the reason why potential matches are often pre-selected, for instance by comparing the parent masses. Spec2Vec in contrast, easily allows to compare large numbers of spectra purely based on their spectral similarity. In addition to the 2 orders of magnitude gain in computation time, our results also suggest that Spec2Vec is a better suited similarity score for detecting related yet different molecules. This can be seen in the considerably lower correlation between high spectral similarity and molecular similarity for the cosine and modified cosine score (fig.3), as well as the observed high fraction of false positives (figure S1).

Unlike in our controlled experiments, we will in practice generally not know if our library contains exact compound matches for our query spectra. We could expect the most reliable library search results when combining different measures. Thus, in the future, we could imagine that Spec2Vec similarity scores together with parent mass matching are very suitable methods for pre-selecting promising candidates for library matching. In a second step, computationally more expensive similarity measures, including cosine and modified cosine scores, but also

scores such as those derived from SIRIUS/CSI:Finger ID[5] could then be added to the Spec2Vec similarities to facilitate a well-informed decision. For instance, high Spec2Vec similarity combined with low cosine score could be used as a signature of a chemically highly related yet distinct compound. Having both a very high Spec2vec and cosine score together with a parent mass match would then suggest an exact match.

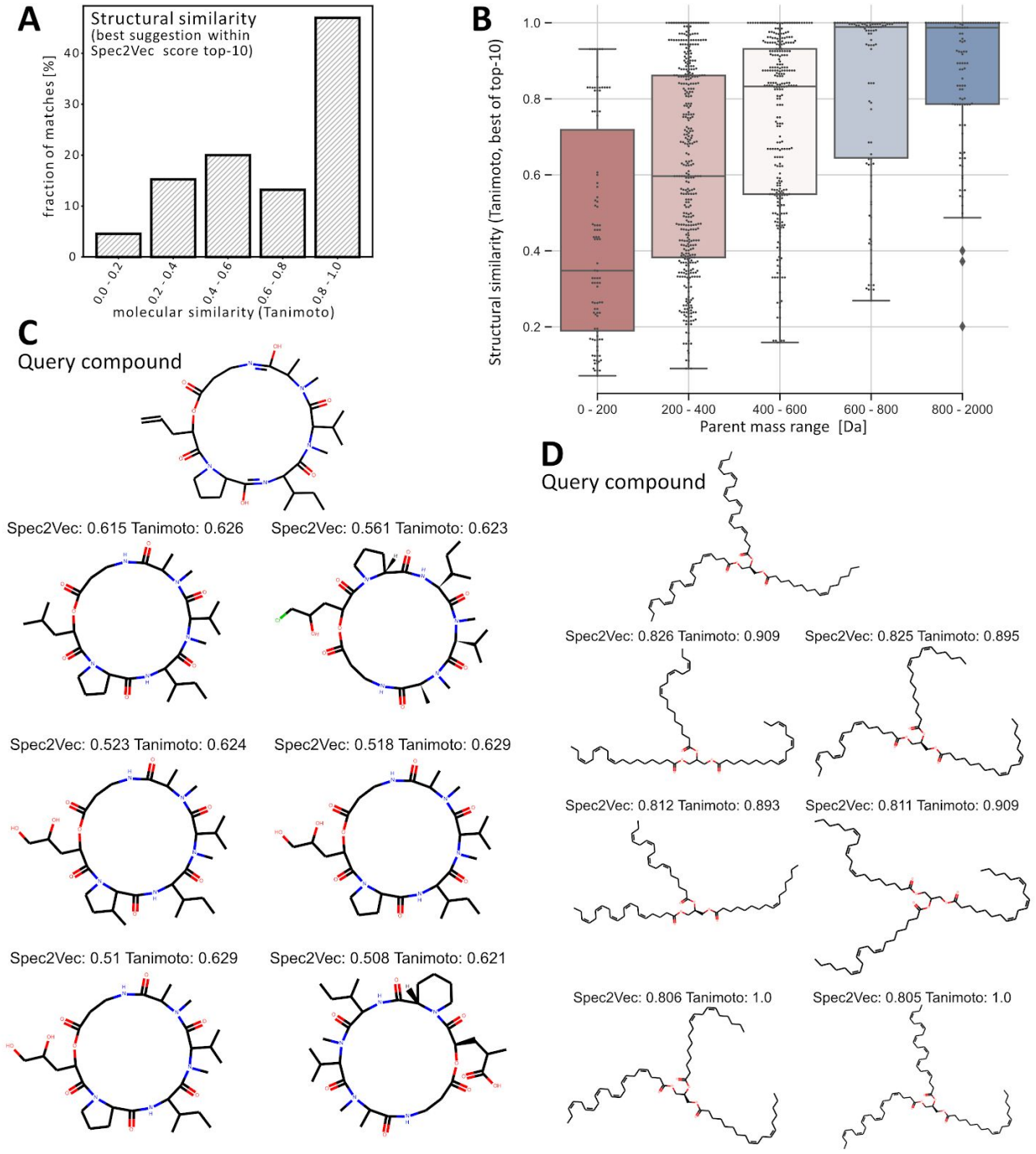



**Figure 5**. Matching of unknown molecules (not part of library) using Spec2vecs similarities. All spectra of 200 randomly selected Inchikeys (1030 spectra) were removed from the **AllPositive** dataset. Using a word2vec model that was trained on the remaining dataset, also excluding non-annotated spectra (76,062 spectra), each removed "query" spectrum was compared to the dataset by only using the Spec2Vec similarity score. (A) shows a histogram of the best structural similarity score out of the

found top-10 Spec2Vec similarities for each query. For nearly 6 out of 10 queries, Spec2Vec finds a match with a structural similarity score > 0.6 reflecting high molecular similarity. (B) The quality of the suggested matches is highly dependent on the mass of the query compound. In particular for larger molecules (> 400 Da), Spec2Vec similarities allow finding highly similar molecules. (C+D) Examples of unknown molecules (not part of library) that are compared to all library spectra to find most similar matches using Spec2Vec. In both cases the algorithm is able to return very related molecules to the query molecules that could be used to help with annotating the query spectra or to infer its chemical class.

## Network analysis.

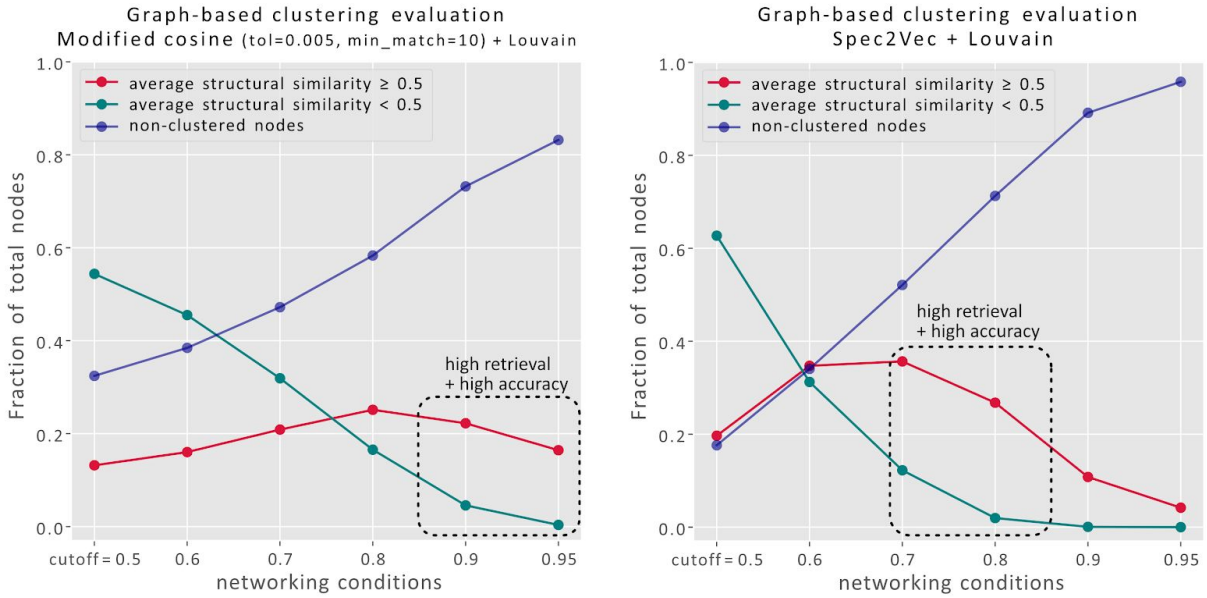



**Figure 6.** The better correlation between Spec2Vec spectral similarity with the actual structural similarity also translates into a more coherent clustering ability. Clustering is done here by creating edges between spectra (=nodes) for similarities above a certain cutoff (max. 10 links per node). To make the resulting clustering more robust and better comparable across different scores, we used the Louvain algorithm to break up the large clusters. Cluster quality here is assessed by measuring the average structural similarity across all linked pairs within each cluster. Setting a structural similarity threshold of 0.5 (see fig. 3A) allows to compare the fraction of spectra that ends up in chemically homogenous clusters (red) with those in more heterogeneous clusters (green) and the fraction of spectra that is not clustered at all (those with no links above set threshold). Dashed squares mark regions of relatively high retrieval (high fraction of clusters with high structural similarity) and high accuracy (large discrepancy between fraction of high structural similarity and low structural similarity clusters). Overall, Spec2Vec allows to cluster higher fractions of spectra into high structural similarity clusters (> 35% of all spectra are in high similarity clusters for a Spec2Vec similarity threshold of 0.7).

Based on the finding that Spec2Vec similarity correlates well with structural similarity (fig.3), we investigated how Spec2Vec could be applied to molecular networking which is becoming an increasingly popular tool for exploring metabolomic datasets [18]. Molecular networking refers to representing spectra as a network in which spectra are nodes connected by edges based upon a user defined cutoff for the similarity score (along with some heuristic pruning). Current molecular networking relies on the modified cosine score [7,18].

Such networks are frequently used to define clusters (or communities) which are termed molecular families. Detecting clusters in complex networks without ground truth is generally

regarded as an ill-defined problem [19], which makes absolute quantitative comparisons difficult. In the present case, clusters depend strongly on the chosen parameters (e.g. similarity cutoff) and algorithm (processing, cleaning and trimming of the network). To better isolate the effect of different similarity scores we have chosen a simple workflow. For each spectrum (=node) the up to 10 highest-scoring links (=edges) to other spectra are added if those links have similarity scores above a given threshold. To improve the overall quality of the clustering, but also to make the results more robust we further apply the Louvain algorithm to split up large, poorly connected clusters [19,20]. Graphs in figure 6 were generated using different similarity thresholds and the resulting clusters were quantified by counting them as well-clustered if the mean structural similarity within all cluster edges was $\geq 0.5$, and poorly-clustered if it was $< 0.5$. The rationale behind setting the cutoff to 0.5 was the low probability of finding scores above 0.5 by chance as observed in the histogram of all possible structural similarity scores within the **UniqueInchikey** dataset (fig. 3A). The results in fig. 6 show that Spec2Vec is able to cluster higher fractions of spectra into high structural similarity clusters when compared to the modified cosine. In terms of computation time, the presented molecular networking procedure requires to calculate the spectral similarities between all possible spectra pairs ($\approx 82 \cdot 10^6$ unique pairs), which is identical to the before described benchmarking. Spec2Vec similarity scores here reduced the computation time by about 2 orders of magnitude. Finally, we expect that Spec2Vec similarity scores bring new options for further improving molecular networking. A simple first proof-of-principle test on combining Spec2Vec similarity and modified cosines scores reveals that combined scores will likely be able to further increase clustering accuracy (see fig. S9). More extensive future work will be necessary to systematically explore the full potential of such score combinations.

# Discussion

In conclusion, here we introduce a machine learned spectral similarity score with advantageous properties over the currently widely used cosine similarity score (and its popular variant, modified cosine). Inspired by natural language processing, much like topic modelling enabled substructure finding in metabolomics data[21], we here show how a text mining algorithm Word2Vec[17] can find meaningful relations between mass fragments and neutral losses in mass fragmentation spectra. We demonstrate how the Spec2Vec score better resembles the structural similarity of fragmented molecules and outperforms the cosine score in key tasks underpinning metabolomics analyses including library matching. One limitation of Spec2Vec as compared to cosine scores is that it needs training data to learn the fragment peak relationships; however, since this not necessarily needs to be library spectra and in light of the enormous increase in publicly available metabolomics data, we do not see this as a major bottleneck. Our use of library data in this study was purely so that spectral similarities could be compared with the similarities of the known structures. Training time is not a limitation:

training the embedding on 95,320 spectra took 40 minutes (when training for 15 iterations) on a standard CPU (here: Intel i7-8550U).

Spec2Vec is not meant to be the endpoint but rather a start of a new direction in spectral similarity scores. The low computational costs of Spec2Vec similarity scores make it possible to run extensive searches on very large library datasets. This makes it particularly suited to act as a pre-selection funnel that can easily be extended with a set of computationally more expensive similarity measures in the future. One could think of adding relevant mass differences as input to train the model, for example following the approach of Kreitzberg et al.[22].

Finally, metabolomics is increasingly used as a tool to understand metabolic profiles and to perform integrative systems biology; furthermore, the use of MS/MS data has been promoted by community based platforms such as MassBank[23], MS2LDA[21] and GNPS[18]. We would like the community to use our novel score and therefore we created the modular matchms[24] and Spec2Vec[25] packages that can easily be incorporated in pipelines such as GNPS. It could also be used in other systems that rely upon spectral similarity. For example, SIRIUS and IOKR both use kernel matrices that are constructed from spectral similarities and whether Spec2Vec similarities could improve these pipelines is clearly worth investigation. We have also only demonstrated performance on LC-MS data. It will also be interesting to assess the utility of Spec2Vec for GC-MS data. When measuring molecules with GC-MS, precursor m/z values are usually not measured. This means that the initial precursor filtering to reduce the number of similarity calculations that can be done for LC-MS is not possible suggesting that the time savings available with Spec2Vec would be particularly desirable. Our work represents the first machine learning inspired spectral similarity score and, given its central place in metabolomics analyses, we believe Spec2Vec opens up new possibilities that will impact metabolomics analyses across all disciplines including clinical, food, and microbial metabolomics as well as biomarker and natural products discovery by improving identification, annotation, and networking.

# Methods

## Data preparation

The current study was done using a large LC-MS dataset provided on GNPS containing 154,820 spectra (https://gnps-external.ucsd.edu/gnpslibrary/ALL_GNPS.json from 2020-05-11). The provided metadata was cleaned and corrected using matchms[24] resulting in 94,121 spectra with InchiKey annotations. Using key metadata information such as compound names, chemical formulas and parent masses an extensive automated lookup search was run against PubChem[26] (via pubchempy[27]). As a result, 128,042 out of 154,820 spectra could be linked to an InchiKey (14,978 unique InchiKeys in the first 14 characters).

The here used **AllPositive** subset contains all spectra with positive ionization mode containing 112,956 spectra, out of which 92,954 with InchiKey (13,505 spectra with unique planar InchiKeys in first 14 characters before Spec2Vec related filtering).
We also worked with the considerably smaller subset **UniqueInchiKeys** which was reduced on purpose to be accessible for extensive benchmarking. It contains only one spectrum for every unique InchiKey from the AllPositive dataset (see supplemental or notebooks for details on the selection procedure).
In a next step we removed all peaks with m/z ratios outside the range [0, 1000] and discarded all spectra with less than 10 peaks. This left us with 95,320 spectra (out of which 77,092 with InchiKey) for the **AllPositive** dataset, and 12,797 spectra (all with InchiKey) for the **UniqueInchiKeys** dataset.
Since Spec2Vec similarity scores are conceptually very different from cosine-like similarity scores we decided to use two different peak filtering procedures for the two methods. For both the cosine and modified cosine score calculations we ignored all peaks with relative intensities < 0.01 compared the highest intensity peak. This is both to remove potential noise, but also to reduce the computational costs for the classical similarity scores.
Spec2vec is comparing spectrum documents using language model analogies. For the underlying word2vec models we hence aimed at training on documents of comparable size which is achieved by removing excessive amounts of low intensity peaks. Since we expect that larger molecules on average will produce a higher number of meaningful fragmentation peaks, the maximum number of kept peaks per spectrum was set to scale linearly with the parent mass:

$$max(n_{peaks}) = 0.5 \cdot parentmass\,.$$

To assess if this procedure of having different peak filtering for the different similarity scores was indeed doing justice to the cosine and modified cosine score, we also repeated the library matching with cosine scores computed based on the Spec2Vec-processed data. This resulted in a slightly lower overall performance of the cosine score based library matching.

## From spectrum to document

After processing, spectra are converted to documents. For this, every peak is represented by a word that contains it's position up to a defined decimal precision ("peak@xxx.xx"). For all presented results a binning of two decimals was used, so that a peak at m/z 200.445 translates into the word "peak@200.45". In addition to all peaks of a spectrum, neutral losses between 5.0 and 200.0 Da were added as "loss@xxx.xx". Neutral losses are calculated as $precursor_{m/z} - peak_{m/z}$. A list of all created peak and loss words is what we here refer to as a document.
A word2vec [17] model is trained on all documents of a chosen dataset using gensim [28]. However, Spec2Vec in several aspects differs significantly from typical NLP applications so that some key

parameters of the model also differ notably from the default settings. First of all, peaks in the mass spectra have no particular order that is comparable to the word order in a document. We hence set the window-size to 500, which in our case means that the entire spectrum (i.e. the entire document) counts as context. The two key types of word2vec models are skip gram and CBOW, the latter was generally observed to perform better in our case.

Although it is often considered that longer training will improve the results of a word2vec model, we found that this does not necessarily hold for our Spec2Vec spectral similarity measures. When using negative sampling during the training, model performance was observed to break down after long training runs. At the same time, however, including negative sampling allowed to obtain better overall results. Generally we found that training a model with negative sampling (negative=5) and 15 (**AllPositive**) up to 50 (**UniqueInchikey**) epochs were best suited for obtaining close to optimal model performance (see supplemental materials and fig. S4).

## Spec2Vec similarity score

Spec2Vec similarity scores are derived on the basis of a pre-trained word2vec model. Word2Vec learns relationships between words (=peaks/losses) from co-occurrences across the seen documents. It then allows to represent words by abstract word vectors (so called word embeddings) in such a way that words of similar meaning are placed close to each other. With Spec2Vec our main interest lies in comparing entire spectra, which can be described by the sum of their words. To account for the dependency between peak relevance and peak intensity we calculate a spectrum vector $v_S$ as a weighted sum:

$$v_S = \sum_{i=1}^{n} \sqrt{w_i} \cdot v_i \ ,$$

with $w_i$ the intensity (normalized to maximum intensity=1) and $v_i$ the word vector of peak $i$.

For the similarity between two spectra we then compute the cosine score between two spectrum vectors:

$$g(s_1, s_2) \ = \ cosine(v_{S1}, \ v_{S2})$$

In practice we expect that Spec2Vec similarities will be most interesting to use when the underlying word2vec was trained on a large reference dataset containing many different fragments and losses and their structural relationships. It can also mean that it will be applied to spectra that were not part of the training data. In those cases some words (=peaks) of a given spectra might be unknown to the model. In those instances we can estimate the impact of the missing words by assessing the uncovered weighted part of a spectrum:

$$missing\ fraction = 1 \ - \frac{\sum\limits_{i,\, w_i \varepsilon model} \sqrt{w_i}}{\sum\limits_{i=1}^{n} \sqrt{w_i}} \ .$$

Having few unknown peaks of low intensity in a spectrum will count only little to the missing fraction, whereas high numbers of unknown peaks or few unknown peaks of high intensity will

result in a high missing fraction. By setting a threshold for the missing fraction (e.g. <0.05), returning Spec2Vec similarity scores for spectra far outside the learned peaks (and losses) can be avoided.

### Structural similarity

Assessing the structural similarity between two molecules remains a complex topic. Finding the best measure to define structural similarity between two molecules lies outside the scope of this study; most recent studies converge to the Tanimoto similarity as one of the most practical and well-performing measure[29]. Thus, for the presented results, the structural similarity was measured by taking the Tanimoto similarity (Jaccard index) based on daylight-like molecular fingerprints (rdkit molecular fingerprints, version 2020.03.2, 2048 bits, derived using rdkit [30] via matchms[24]).

## Acknowledgement

We want to thank Marnix Medema, Stefan Verhoeven, Michelle Schorn for many helpful discussions. Further, we thank the eScience Center generalization team, and notably Jurriaan Spaaks and Faruk Diblen for their help and input around code design, packaging, and testing. J.J.J.v.d.H. acknowledges funding from an ASDI eScience grant, ASDI.2017.030, from the Netherlands eScience Center—NLeSC.

## Code, trained models, and data

The underlying code was developed into two Python packages to handle and compare mass spectra, matchms (https://github.com/matchms/matchms) and spec2vec (https://github.com/iomega/spec2vec). Both packages are freely available and can be installed via conda.

All additional functions to analyse the data and to create the presented plots can be found under https://github.com/iomega/spec2vec_gnps_data_analysis. This repository also contains extensive Jupyter notebooks to document the entire workflow from raw data to the figures presented in this work.

The two most important trained Word2Vec models used in this work can be downloaded from https://zenodo.org/record/3978054 (trained on UniqueInchikey dataset) and https://zenodo.org/record/3978070 (trained on AllPositive dataset).

The pre-processed, cleaned dataset with all positive ionization mode spectra can be downloaded from https://zenodo.org/record/3978118, the original raw data can be accessed from https://zenodo.org/record/3979010.

Calculated all-vs-all similarity score matrices for cosine score, modified cosine score, and fingerprint-based similarity (Tanimoto) for the UniqueInchikey dataset can be found on https://zenodo.org/record/3979074.

## References


1. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **13**, 263–269 (2012).
2. May, M. Big data, big picture: Metabolomics meets systems biology. *Science* **356**, 646–648 (2017).
3. Fu, J. *et al.* System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nat. Genet.* **41**, 166–167 (2009).
4. Spicer, R., Salek, R. M., Moreno, P., Cañueto, D. & Steinbeck, C. Navigating freely-available software tools for metabolomics analysis. *Metabolomics* **13**, 106 (2017).
5. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 12580–12585 (2015).
6. Blaženović, I., Kind, T., Ji, J. & Fiehn, O. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **8**, (2018).
7. Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E1743-1752 (2012).
8. Demuth, W., Karlovits, M. & Varmuza, K. Spectral similarity versus structural similarity: mass spectrometry. *Anal. Chim. Acta* **516**, 75–85 (2004).
9. Schollée, J. E. *et al.* Similarity of High-Resolution Tandem Mass Spectrometry Spectra of Structurally Related Micropollutants and Transformation Products. *J. Am. Soc. Mass*

*Spectrom.* **28**, 2692–2704 (2017).

10. Aguilar-Mogas, A., Sales-Pardo, M., Navarro, M., Guimerà, R. & Yanes, O. iMet: A Network-Based Computational Tool To Assist in the Annotation of Metabolites from Tandem Mass Spectra. *Anal. Chem.* **89**, 3474–3482 (2017).

11. Schymanski, E. L. *et al.* Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).

12. Ridder, L., van der Hooft, J. J. J. & Verhoeven, S. Automatic Compound Annotation from Mass Spectrometry Data Using MAGMa. *Mass Spectrom.* **3**, (2014).

13. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).

14. Brouard, C. *et al.* Fast metabolite identification with Input Output Kernel Regression. *Bioinformatics* **32**, i28–i36 (2016).

15. Ji, H., Xu, Y., Lu, H. & Zhang, Z. Deep MS/MS-Aided Structural-Similarity Scoring for Unknown Metabolite Identification. *Anal. Chem.* **91**, 5629–5637 (2019).

16. Shen, X. *et al.* Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat. Commun.* **10**, 1516 (2019).

17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. in *Advances in neural information processing systems* 3111–3119 (2013).

18. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with GNPS. *Nat. Biotechnol.* **34**, 828–837 (2016).

19. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Phys. Rep.* **659**, 1–44 (2016).

20. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

21. Hooft, J. J. J. van der, Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci.* **113**, 13738–13743 (2016).

22. Kreitzberg, P. A., Bern, M., Shu, Q., Yang, F. & Serang, O. Alphabet Projection of Spectra. *J. Proteome Res.* **18**, 3268–3281 (2019).

23. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).

24. Huber, F. *et al.* matchms - processing and similarity evaluation of mass spectrometry data. *bioRxiv* 2020.08.06.239244 (2020) doi:10.1101/2020.08.06.239244.

25. Huber, F. *et al. spec2vec*. (Zenodo, 2020). doi:10.5281/zenodo.3873169.

26. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).

27. Matt Swain *et al. mcs07/PubChemPy: PubChemPy v1.0.4*. (Zenodo, 2017). doi:10.5281/zenodo.541438.

28. Řehůřek, R. & Sojka, P. Software Framework for Topic Modelling with Large Corpora. in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* 45–50 (ELRA, 2010).

29. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminformatics* **7**, 20 (2015).

30. Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org.

# Supplemental material

## Data processing protocol

**Data pre-processing:**

1. Import all spectra from GNPS website via matchms json import (all spectra file from https://gnps-external.ucsd.edu/gnpslibrary/ALL_GNPS.json from 2020-05-11)
   Found non-empy spectra: 154820 spectra
2. Run default matchms filters to clean, correct, and infer missing metadata. Results in:
   94462 spectra with InChI (15822 unique)
   94155 spectra with Smiles (20542 unique)
   94121 spectra with InchiKey (13505 unique in first 14 characters)
3. Run extensive automated PubChem lookup (code provided in manuscript repository). Results in:
   128103 spectra with InChI (17620 unique)
   128052 spectra with Smiles (23097 unique)
   128042 spectra with InchiKey (14978 unique in first 14 characters)

**Creation of subsets:**

Creation of 4 different datasets:

- **AllGNPS:** All spectra after matchms cleaning and PubChem lookups (154820 spectra)
- **AllPositive:** all spectra from AllGNPS dataset with positive ionization mode.
  Total: 112956 spectra
  92997 spectra with InChI (16071 unique)
  92964 spectra with Smiles (20540 unique)
  92954 spectra with InchiKey (13717 unique in first 14 characters)
- **AllPositiveAnnotated**
- **UniqueInchiKeys:** Reduced dataset used for benchmarking. Keep only one spectrum for every unique InchiKey from the AllPositive dataset. Spectra are selected by:
  i) If possible, select spectra with > 10 peaks above 1% of maximum peak intensity
  ii) Out of those (if multiple): select best library quality level: 1> 2> 3
  iii) Out of those (if multiple): select spectrum with most peaks above 1% of maximum peak intensity
  iv) if still multiple, pick random!
  This gives 13717 spectra, out of which 12797 spectra contain >= 10 peaks.

**Data post-processing for "classical" similarity scores (cosine, modified cosine)**

1. Remove peaks with m/z outside [0, 1000]
2. Remove spectra with < 10 peaks remaining
3. Remove peaks with intensities < 0.01 maximum peak intensity (smaller peaks will slow down score calculation, but won't contribute much to the overall scores)

**Data post-processing for spec2vec**

1. Remove peaks with m/z outside [0, 1000]
2. Remove spectra with < 10 peaks remaining
3. Reduce number of peaks using matchms "reduce_number_of_peaks" filter.
   Parameters: n_required = 10, ratio_desired = 0.5
   Spec2vec is comparing spectrum documents using language model analogies. For the underlying word2vec models we aimed at training on documents of roughly comparable size to ensure that spectra will also get comparable attention during model training. The raw data contained spectra with numbers of peaks ranging from 10 (our own set threshold to include spectra) up to several 10,000s of peaks. We hence removed excessive amounts of low intensity peaks. To account for the fact that larger molecules on average show more fragmentation peaks, the maximum number of kept peaks per spectrum was set to scale linearly with the parent mass:

$$max(n_{peaks}) = 0.5 \cdot parentmass\,.$$

## Analysis of modified cosine reliability

We noted a large amount of false positive among the spectra pairs with high cosine or modified cosine scores. With false positives we here mean spectral pairs that receive a high spectra similarity but a very low structural similarity score. This clearly shows in histograms of all structural similarities which belonged to pairs that received spectra similarity scores above a set threshold. For a given score, for instance modified cosine, we would select all pairs across the **UniqueInchikey** dataset (12,797 spectra, total of 81,875,206 unique pairs when excluding pairs of spectra with themselves) with a modified cosine score > 0.7, 0.8, etc. and create a histogram of the corresponding structural similarity values (fig. S1). Even though we already use a relatively small peak m/z tolerance of 0.005 Da (preventing the collapse of too many mass fragments with different elemental formulas into one mass bin), the *min_match* criteria also clearly has to be raised to obtain acceptable results. In comparison, Spec2Vec similarities show a visibly lower amount of false positives (Fig. S1, lower right).
We also consider this one of the reasons for the notably poorer correlation between cosine and modified cosines scores and structural similarities (Figure 3 in main manuscript).

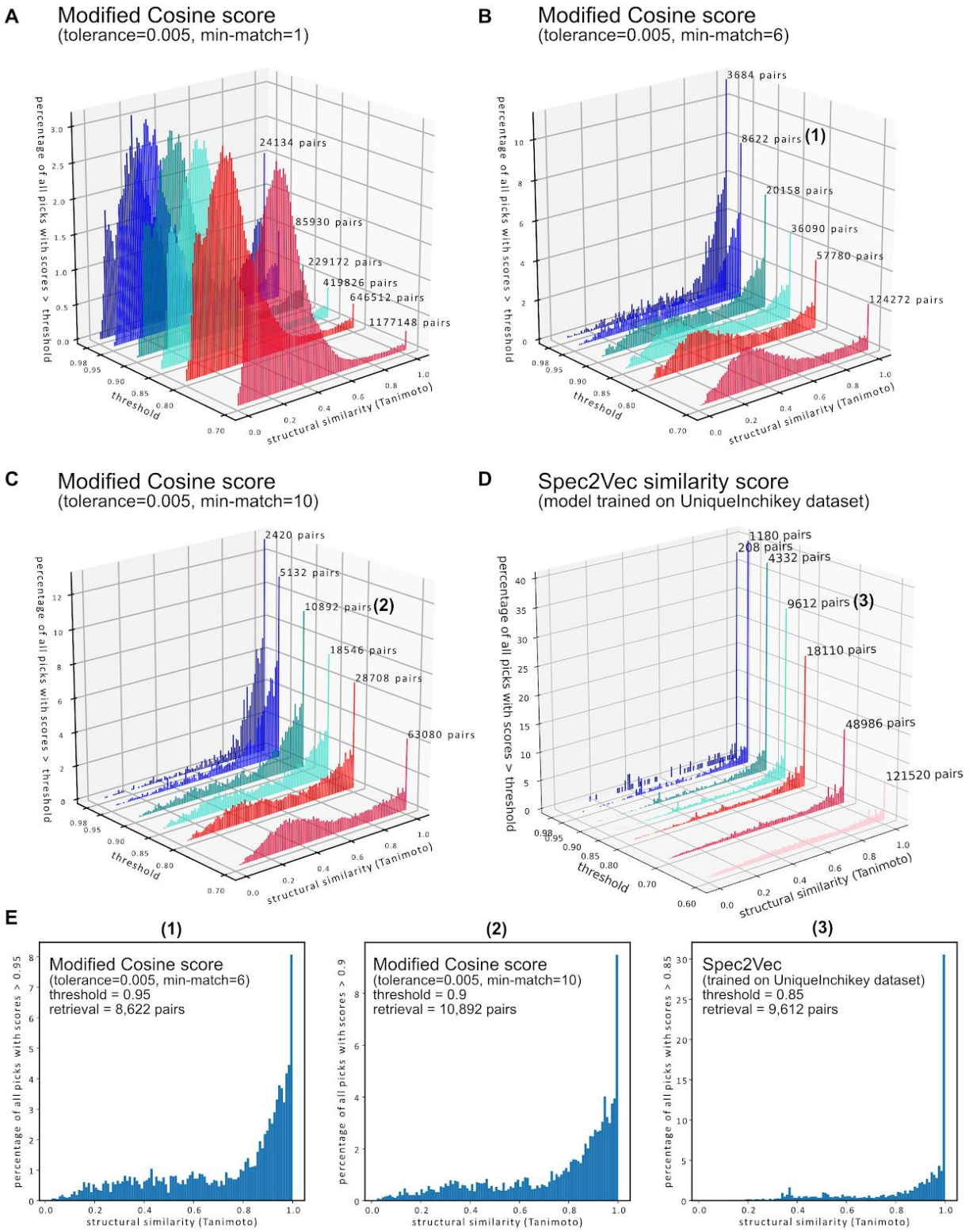



**Figure S1**. Histograms of all structural similarity scores (Tanimoto) within the UniqueInchikey dataset for which the respective similarity scores are above a set threshold. Scores between identical spectra are excluded here. Thresholds between 0.7 to 0.98 were used to create the displayed histograms and are shown on the plot axis. The number of pairs that fitted the criteria (min-match and > threshold) is written on the top right of each histogram. (A) Modified cosine scores without requiring multiple machine peaks between two spectra results in drastic numbers of false positives (=low Tanimoto scores despite

spectra similarity scores > threshold). (B) Raising the minimum number of matches to 6 already improves the reliability of the modified cosine score notably. Still, even for comparably high thresholds (scores > 0.9 or > 0.95) a considerable fraction of all pairs with high modified cosine scores do not correspond to similar molecules. (C) For a minimum of 10 matching peaks the reliability further improved, but also the number of pairs fulfilling this criteria continues to drop. (D) Spec2Vec similarities also show notable levels of false positives, in particular for lower thresholds (0.7), but generally are visibly more shifted towards high Tanimoto scores when compared to the modified Cosine scores. Absolute threshold values are hard to compare. For instance, a Spec2Vec similarity of 0.9 could be more or less common than a Cosine score of 0.9 for a given setting. In (E) we hence compare three histograms that represent roughly equal numbers of spectra pairs (between 8,622 and 10,892). This shows that for comparable retrieval rates, high Spec2Vec scores correlate less frequently with very low molecular similarities.

## Benchmarking and parameter tuning for the cosines score

Cosine scores were calculated for all-vs-all spectra of the **UniqueInchikey** dataset using different tolerances and minimum matching peak thresholds. This required computing similarity scores for all possible pairs between the 12,797 spectra (81,875,206 unique pairs), which --depending on the pre-processing and parameters used-- takes several hours on a standard CPU to compute. The precise computation time is highly dependent on the peak filtering, for the present settings, run on a Intel i7-8550U, the computation took about 6 hours. The number of matching peaks and the actual scores can be computed simultaneously (using matchms[24]).

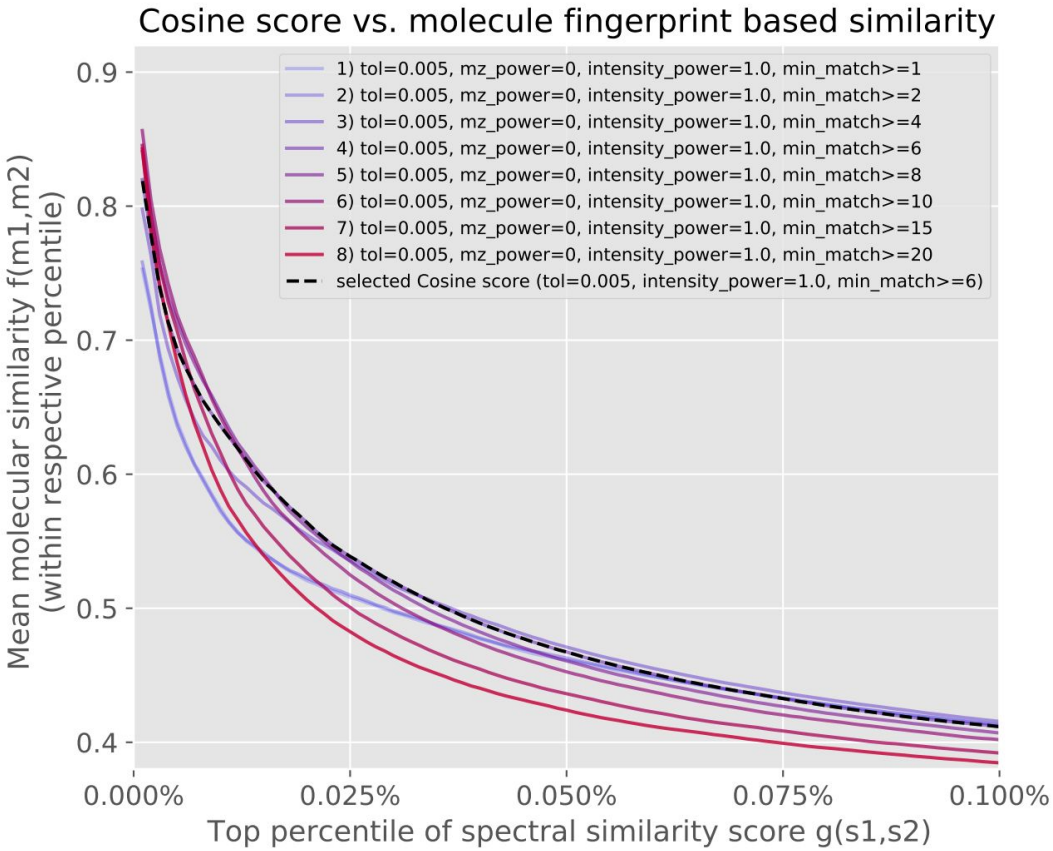



**Figure S2**. Parameter search for cosine similarity score. Varying from 1 to 20 the minimum number of matching peaks necessary to calculate a cosine score. For all following comparisons, min_match=6 was chosen.

Cosine scores come in many flavors, which makes it impossible to systematically compare all possible parameters settings and implementations. Apart from what we would consider the most basic implementation used here, we also tested cosine scores that weigh peaks according to their m/z ratio or their relative intensity. Settings that were tested in-depth stem from a suggestion from Demuth et al.[8] , as well as from current implementation used in NIST and Massbank[9].

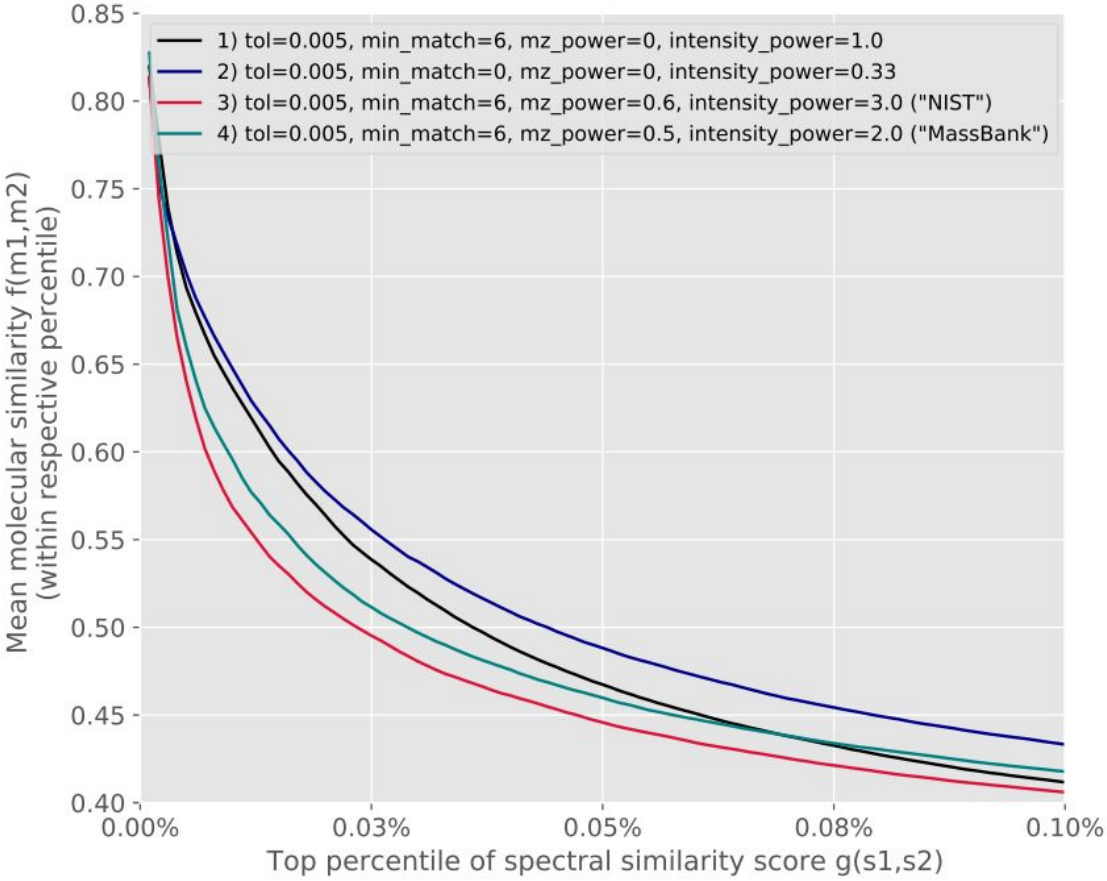



**Figure S3**. Comparing different Cosine score flavors. Four different mz_power/intensity_power parameter settings were tested. The plot shows the respective results with the best performing min_match criterion.

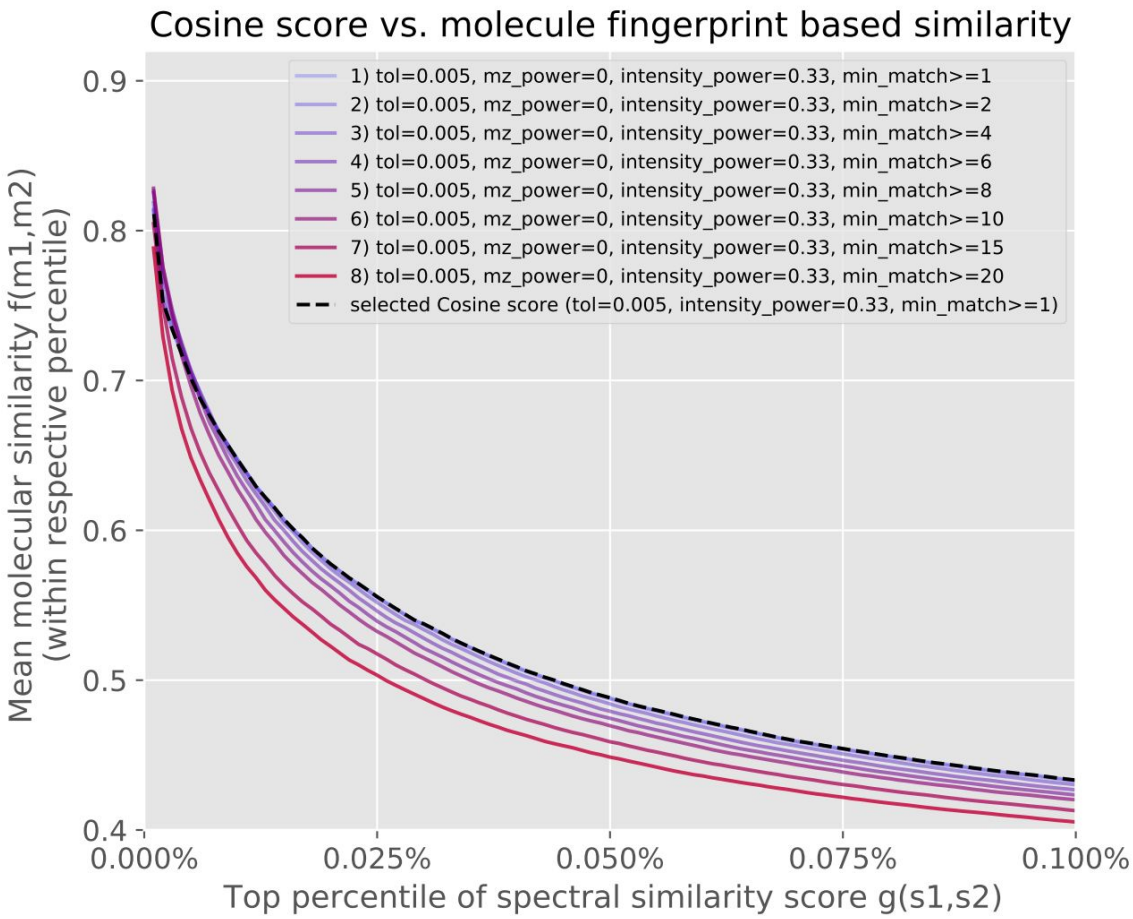



**Figure S4**. Benchmarking of cosine score with intensity_power=0.33 (Demuth et al. [8]) for different min_match settings.

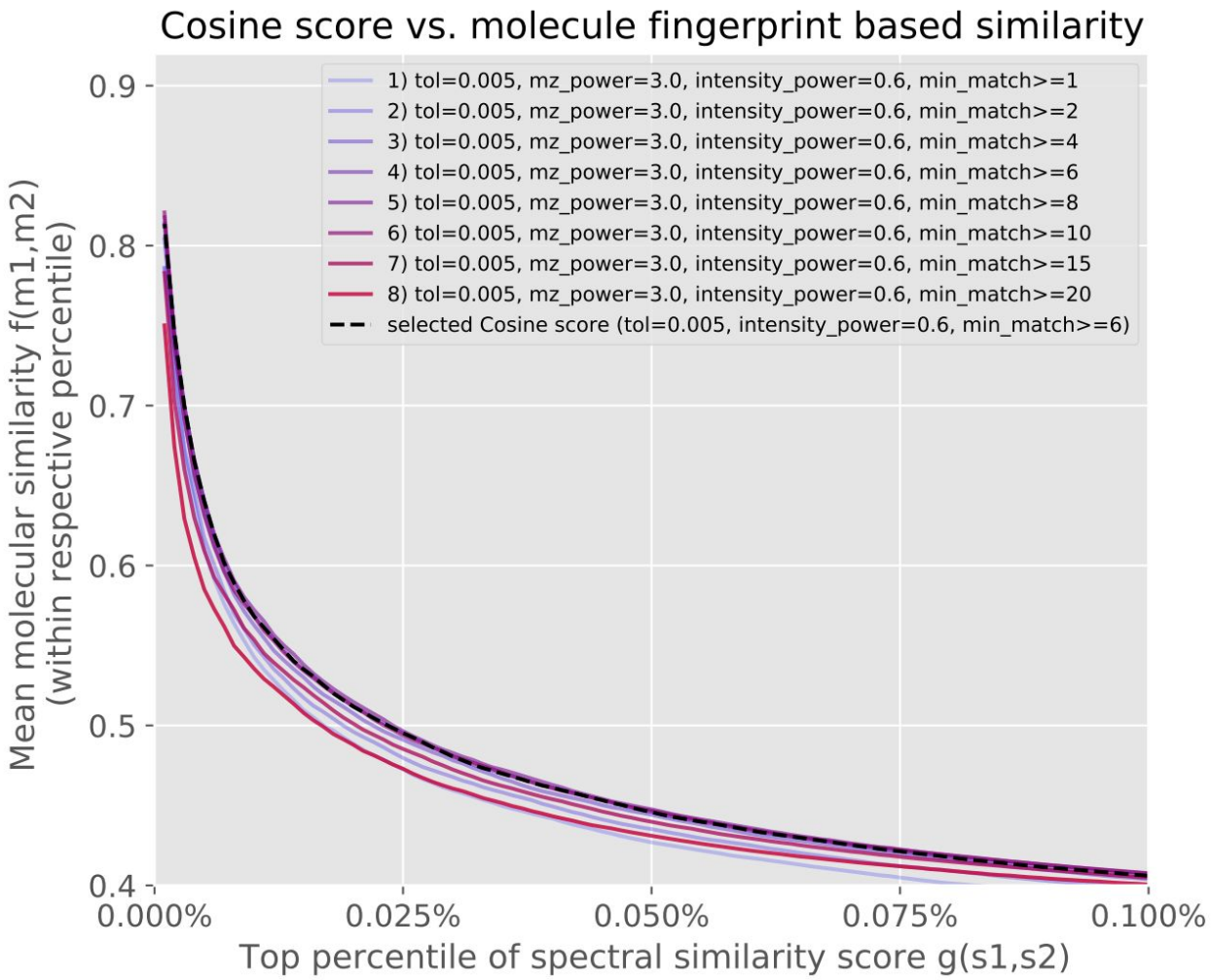



**Figure S5**. Benchmarking of cosine score with mz_power=3.0 and intensity_power=0.6 (NIST settings [9]) for different min_match settings.

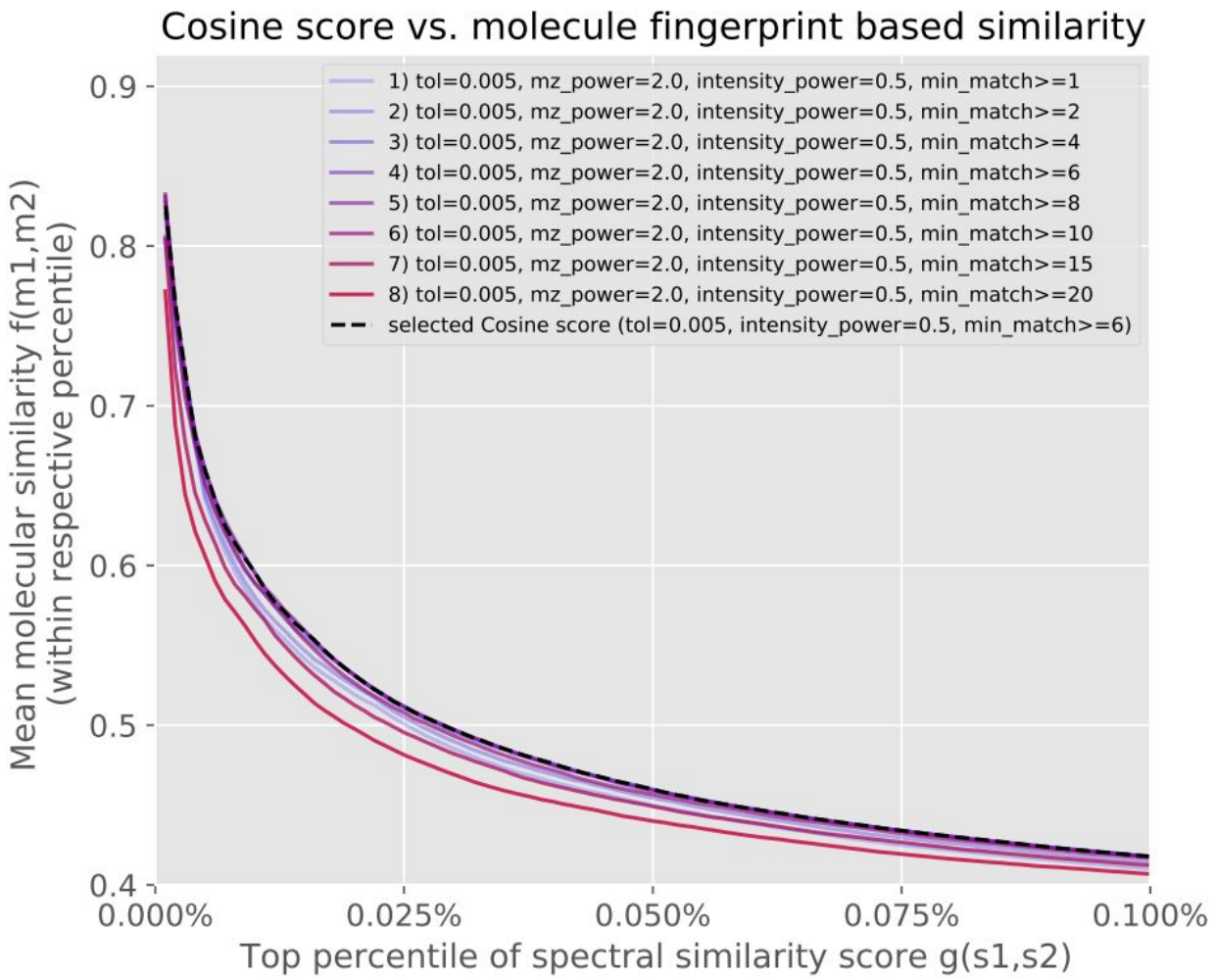
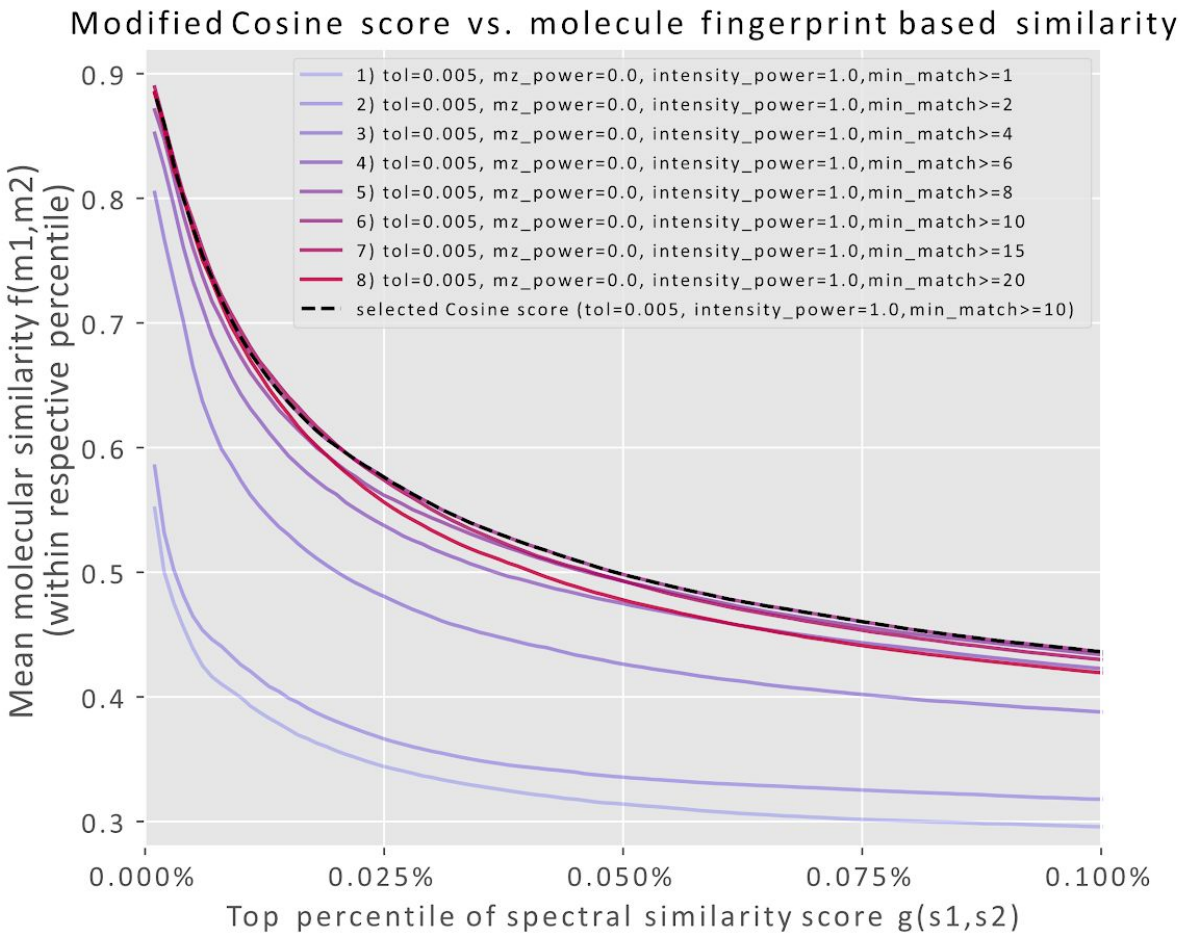



**Figure S6**. Benchmarking of cosine score with mz_power=2.0 and intensity_power=0.5 (MassBank settings [9]) for different min_match settings.

## Benchmarking and parameter tuning for the modified cosines score

As for the cosine score, modified cosine scores were calculated for all-vs-all spectra of the **UniqueInchikeys** dataset using a tolerance of 0.005 Da together with different minimum matching peak thresholds (fig. S7).



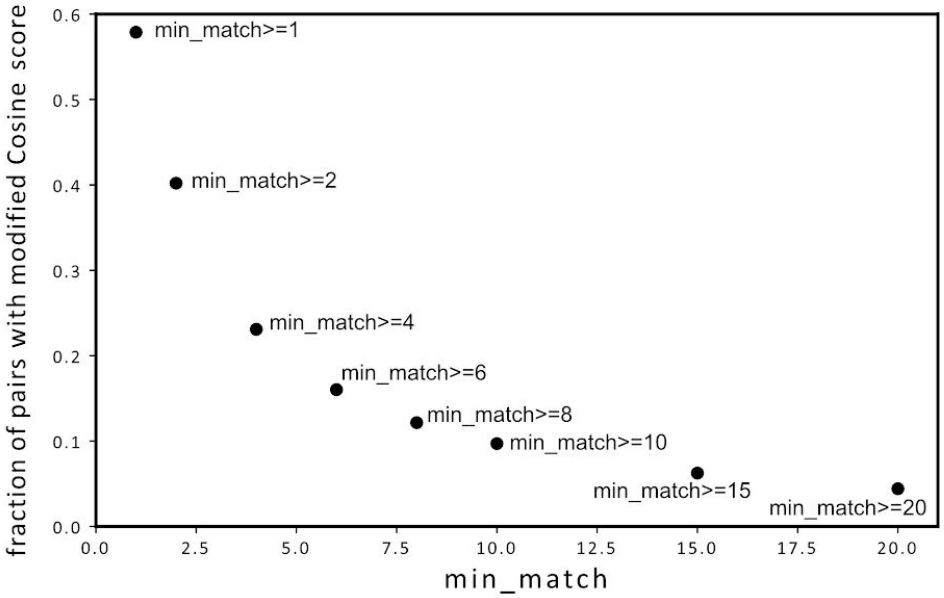




**Figure S7**. Parameter search for modified cosine similarity score. Varying from 1 to 20 the minimum number of matching peaks necessary to calculate a cosine score. For all following comparisons, min_match=10 was chosen. Please note that with

increasing min_match parameter, more and more spectra pairs will not receive a Cosine score (bottom plot). For min_match=10, for instance, less than 10% of all spectra pairs will receive a Cosine score. For min_match=20 this drops below 5%. This also means potentially losing correctly matching pairs that just happen to not have sufficient shared peaks (unlike for Spec2Vec which would return a score for any pair).mat

## Spec2Vec model parameters

The underlying word2vec models were trained using the spec2vec Python library running Gensim. The key model parameters were

- Window size = 500
- Word vector dimension = 300
- Negative sampling
- CBOW mode
- Initial learning rate = 0.025
- Learning rate decay per iteration (per epoch) = 0.00025

We compared models after different iterations using the benchmarking dataset **UniqueInchikey** to find suitable numbers of training iterations (fig. S8). Generally training was comparably stable over a larger range of iterations. It converges to stable results when switching off negative sampling, but negative sampling resulted in better overall results (not shown here).

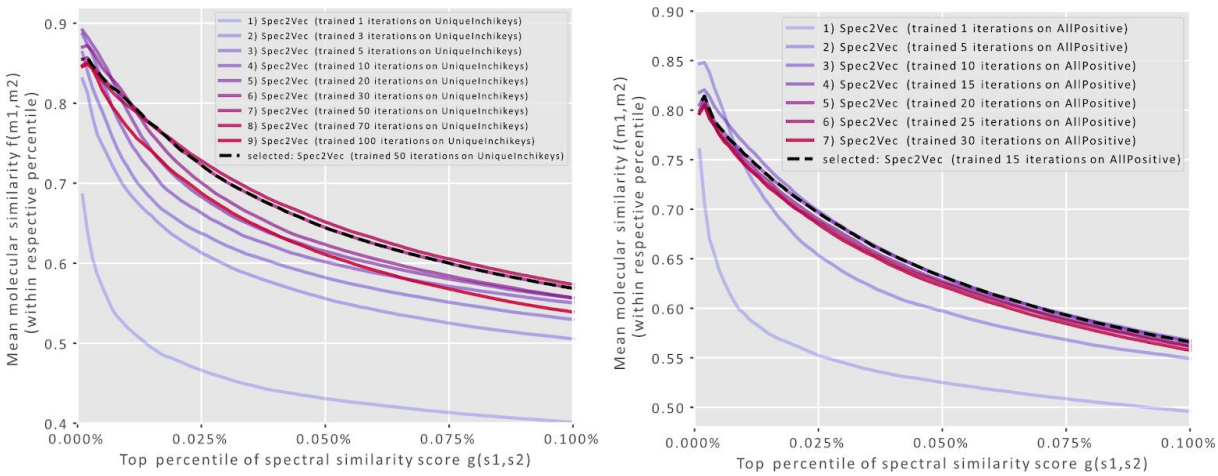



**Figure S8.** Word2Vec models were trained on both the **UniqueInchikey** and the **AllPositive** dataset over many iterations. The overall performance was monitored by the correlation between Spec2Vec similarities and structural similarity and revealed that the model rapidly improved during the first 10-15 epochs (iterations). For the UniqueInchikey dataset we found the best performance around 50 iterations, for the AllPositive dataset we observed that the changes after 15 iterations were rather minor.

## Network analysis

Based on the very different working principles of cosine-based spectral similarity scores, and Spec2Vec similarity, we expect different weaknesses and strengths of both score types. Those differences could be exploited by combining scores. In a very first, and very simplified, test on the described molecular networking task (see main manuscript, fig.6), we observed that a simple linear combination of both modified cosine and Spec2Vec similarities only mildly lowers the total fraction of clustered spectra, but notably reduces fraction of poorly clustered spectra (fig. S9). We can hence increase clustering accuracy by relying on both scores, which seems a promising starting point to build upon in future work.

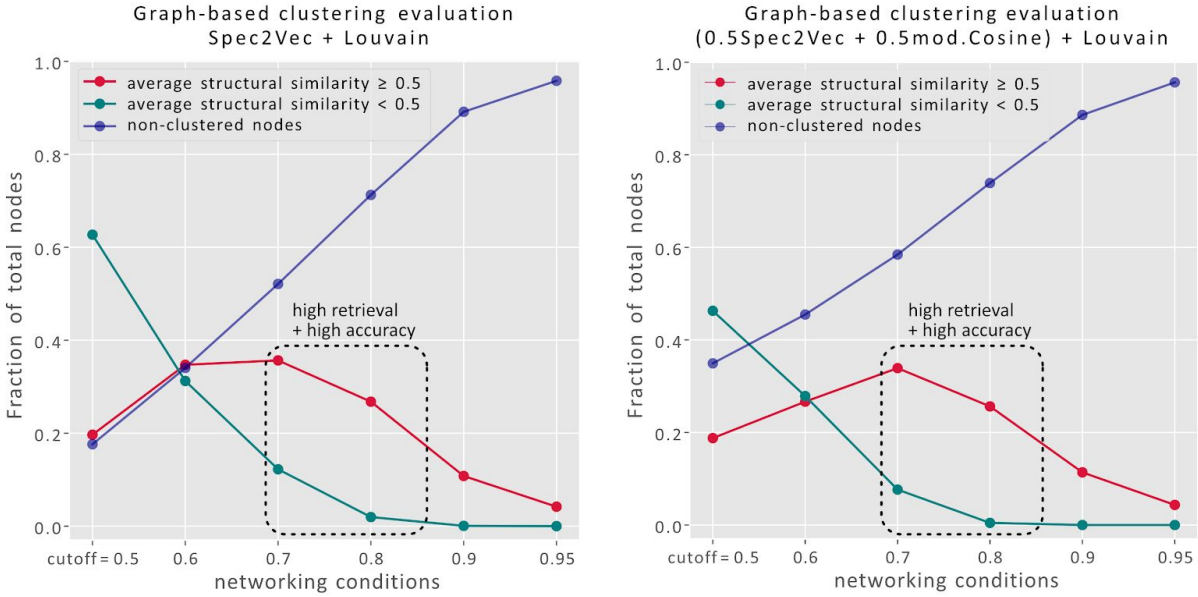



**Figure S9.** Networks were generated from spectra (nodes) by adding links based on spectra similarities, as shown in the main manuscript (see fig. 6). The left plot is identical to the right plot in figure 6 (main manuscript) and was generated by only using Spec2Vec similarity scores. On the right we used a simple combination of Spec2Vec similarities and modified Cosine scores (similarity = 0.5 Spec2Vec similarity + 0.5 mod.Cosine, min_match=10, tolerance=0.005). Already this simple combination leads to a drop of the loss structural similarity clusters (green) with respect to the well-clustered fraction (red), effectively leading to a higher accuracy. We expect that more refined combinations of both similarity scores can be used to further build upon these results.